Volume 14 Number 14 1986 Nucleic Acids Research

Complete nucleotide sequence of the nonstructural protein genes of Semliki Forest virus

Kristiina Takkinen

Recombinant DNA Laboratory, University of Helsinki, Valimotie 7, SF-00380 Helsinki, Finland

Received 15 May 1986; Accepted 14 June 1986

ABSTRACT

The nucleotide sequence coding for the nonstructural proteins of Semliki Forest virus has been determined from cDNA clones. The total length of this region is 7381 nucleotides, it contains an open reading frame starting at position 86 and ending at an UAA stop codon at position 7379-7381. This open reading frame codes for a 2431 amino acids long polyprotein, from which the individual nonstructural proteins are formed by proteolytic processing steps, so that nsP1 is 537, nsP2 798, nsP3 482 and nsP4 614 amino acids. In the closely related Sindbis and Middelburg viruses there is an opal stop codon (UGA) between the genes for nsP3 and nsP4 (1). Interestingly, no stop codon is found in frame in this region of the Semliki Forest virus 42S RNA. In other aspects the amino acid sequence homology between Sindbis, Middelburg and Semliki Forest virus nonstructural proteins is highly significant.

INTRODUCTION

Semliki Forest virus (SFV) belongs to the alphaviruses and is, together with Sindbis virus, the best characterized member of this virus group. The viral nucleocapsid, composed of a capsid protein and a RNA genome, is surrounded by a lipid membrane with viral glycoproteins E1, E2 and E3 (for reviews see 2, 3). The genome of SFV, the 42S RNA, is a single stranded RNA of positive polarity with a 5′ terminal cap structure and poly(A) region at the 3′ end (4, 5, 6). In the early phase of infection the 5′ two thirds of the genome are translated as a polyprotein to the nonstructural (ns) proteins, required for the replication of viral RNA. In the first step of viral replication, a full length minus strand is synthesized, which then serves as a template for new plus 42S RNA synthesis (7, 8). The mRNA for the four viral structural proteins, a subgenomic 26S RNA, is transcribed starting at an internal initiation site on the minus strand 42S RNA template (4, 9). The 26S RNA is also translated as a polyprotein which is proteolytically processed to the four structural proteins. Temperature-sensitive mutants of Sindbis virus, which have defects in the RNA synthesis fall into four complementation groups, suggesting that four different proteins are involved in alphavirus RNA replication (10, 11, 12).

The translation order of the ns proteins of SFV, nsPl (ns70)-nsP2(ns86)-nsP3(ns60)-nsP4(ns72), has been determined by using salt synchronized protein synthesis and specific labeling conditions (14, 15). The present nomenclature of these proteins, nsPl, nsP2, nsP3 and nsP4 reflects to the order of translation (13), the previous ns70, ns86, ns60 and ns72 to their apparent molecular weights. The ns proteins of SFV were first identified in cells infected with a temperature-sensitive mutant, ts-1, which turned out to be an over producer of the ns proteins (16, 17). Various short lived intermediate cleavage products are found in cells infected with ts-mutants. Two intermediate cleavage products of the 250 kd ns polyprotein, ns155 and ns135, are found in cells infected with ts-1 containing the sequences of nsPl plus nsP2 and nsP3 plus nsP4, respectively (18). Another precursor ns220, which is synthesized together with the entire polyprotein ns250 in cells infected with ts-4 or ts-6, is processed to nsPl, nsP2 and nsP3 (19, 20). The cleavage intermediates accumulate in mutant infected cells due to impaired proteolytic processing of the primary translation product (21). All of them can be identified in cells infected with wild type SFV, but in much smaller quantities. The synthesis and processing of the ns proteins of Sindbis virus is different, because of the stop codon between nsP3 and nsP4 (1). The major precursor polyprotein (p230) of Sindbis is processed to nsPl, nsP2 and nsP3. Only minor amounts of the full length ns polyprotein, from which the nsP4 is generated, is synthesized by read through of the stop codon (1, 13).

The proteolytic processing sites of the ns polyprotein of SFV have been determined by direct NH_2-terminal amino acid sequence analysis of radiolabeled ns proteins (22, Kalkkinen et al., to be published). This allows to localize the genes for the individual ns proteins in the nucleotide sequence presented in this paper. The 26S RNA region of SFV has been cloned and sequenced earlier

(23, 24), so now the total structure of the genome of SFV is known. Nucleotide sequence data from other alphaviruses is also wide; at present the complete nucleotide sequence of the genome of Sindbis virus has been determined (13, 25) and parts of the ns genomic region of Middelburg virus has been sequenced (1). This sequence information of the ns protein genes provides possibilities to search for conserved domains of the ns proteins with important functions in the RNA replication of alphaviruses.

MATERIALS AND METHODS

Enzymes and reagents

Enzymes were obtained from commercial sources, isotopes $[\gamma\text{-}^{32}P]$ ATP (3000 Ci/mmol), $[\alpha\text{-}^{32}P]$ dNTP (3000 Ci/mmol), $[\alpha\text{-}^{32}P]$ ddATP (3000 Ci/mmol) and $[\alpha\text{-}^{35}S]$ dATP (~ 600 Ci/mmol) from Amersham. The oligonucleotide primer (15-mer) used for the direct RNA sequencing was synthesized in the Beckman Application Laboratory.

cDNA synthesis and cloning

Viral 42S RNA was isolated from purified virions disrupted with 2 % SDS and was purified on 15-30 % sucrose gradient as described (26). Restriction enzyme fragments of about 100-200 bp in length were isolated as primers for the cDNA synthesis by linear 2.5-10 % polyacrylamide gradient gels (27) from the cDNA clones in pBR322. The desired fragments were eluated by diffusion (28) or electrophoresed onto DEAE-cellulose membranes (NA-45 Schleicher&Schüll) and recovered with 50 mM TRIS-HCl pH 7.5, 1 M NaCl and precipitated with ethanol. 5 µg of viral RNA was ethanol precipitated with 5-10 fold molar excess of primer, washed with 70 % ethanol, dried and dissolved in 5 µl of water. The template-primer mixture was denatured at 90°C for 3 min and annealed at 60°C for 20 min in 50 mM TRIS-HCl pH 8.3, 0.6 M KCl. The reaction mixture (100 µl) for the first strand synthesis contained: 50 mM TRIS-HCl Ph 8.3, 10 mM $MgCl_2$, 140 mM KCl, 5 mM DTT, 500 µM dATP, dGTP, TTP, 250 µM dCTP, 30 µCi of $[\alpha\text{-}^{32}P]$dCTP, 150 U RNase inhibitor and 70-100 U reverse transcriptase (Life Science or Promega Biotec). The synthesis was at 42°C for 60 min, and was stopped by adding EDTA to 20 mM. The cDNA was extracted with phenol and precipitated with ethanol from 2 M ammonium acetate. The second strand synthesis was done either according to Gubler and Hoffman (29) or by the following procedure: The RNA was hydrolyzed at 50°C for 30 min in 65 mM NaOH and the nucleotides were removed by gel filtration on Biogel P-30 (Biorad) and ss cDNA was recovered with ethanol precipitation. The second strand was synthesized with reverse transcriptase as above, omitting RNase inhibitor and $[\alpha\text{-}^{32}P]$dCTP and the concentration of dCTP was 500 μM. The reaction was stopped and cDNA recovered as before. The ds cDNA was then treated with 1 U of S1-nuclease in 50 μl of 50 mM NaCl, 30 mM Na-acetate pH 4.6, 1 mM $ZnCl_2$. The ds cDNA was size selected on a 2 ml column of Sephacryl S-1000 (Pharmacia). The 3′ termini of the cDNA were elongated with dCTP and terminal transferase to give homopolymeric tails (10-30 residues) and vector pBR322, cleaved with PstI, was similarly tailed with dGTP (30), the cDNA was annealed to a molar amount of vector. Some clones were prepared by ligating SalI cleaved cDNA into SalI site of pBR322. Transformation into E. coli HB101 was according to Mandel and Higa (31). Small scale plasmid DNA was prepared (32) from colonies which had the correct antibiotic resistance pattern and were then assayed for colinearity with viral 42S RNA using S1-nuclease test (33) and large scale plasmid DNA was prepared as previously described (33).

DNA sequencing

The nucleotide sequences of the cDNA clones were determined by Maxam and Gilbert (28) or Sanger dideoxy (34) methods. The restriction enzyme maps of the inserts were constructed by partial digestions of end labeled insert fragments (35) or double digestions. For chemical sequencing the 5′ ends of the fragments were first dephosphorylated with calf intestinal alkaline phosphatase and then labeled with T4-polynucleotide kinase and $[\gamma\text{-}^{32}P]$ATP (27). The 3′ ends were labeled either with Klenow fragment and appropriate $[\alpha\text{-}^{32}P]$dNTP (36) or by terminal transferase and $[\alpha\text{-}^{32}P]$ddATP (30). Labeled DNA fragments were digested with secondary restriction enzyme and isolated from polyacrylamide gradient gels or strand separated (28) and purified from contaminating polyacrylamide by mini DE 52 cellulose columns (37). DNA sequences were analyzed on 20 %, 8 % and 6 % polyacrylamide gels (38).

For the dideoxy sequencing, insert specific restriction fragments were isolated by polyacrylamide gradient gels, ligated to

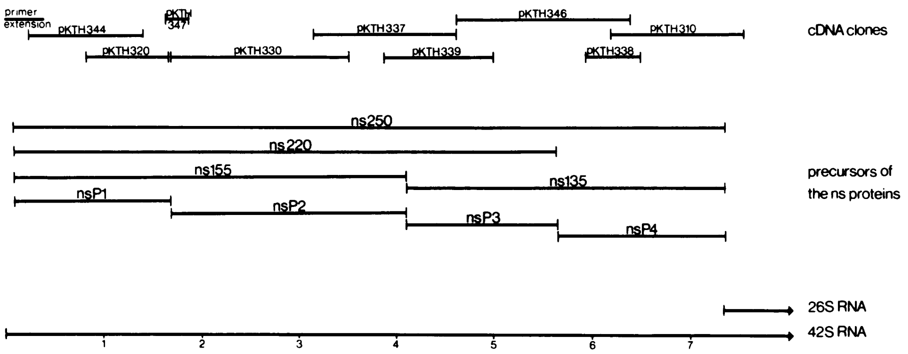

Figure 1 Schematic representation of the nine cDNA clones in the ns protein coding region of SFV used in the nucleotide sequence determination. The nucleotide sequence of the 5′ end of the genome has been determined earlier by primer extension (37). The ns polyprotein, the intermediate and the final proteolytic cleavage products are in the middle. The bottom line indicates the scale in kilobases and shows the location of the 5′ end of 26S RNA (45).

dephosphorylated M13 vectors, and transfected to JM103 (39). The sequences from recombinants with complementary strands were determined (39) using $[\alpha\text{-}^{35}S]$ dATP as the label (40).
For the direct RNA sequencing about 20 ng of the primer was annealed with 1-2 µg of viral RNA and sequenced with reverse transcriptase (41) using $[\alpha\text{-}^{35}S]$ dATP as the label.

RESULTS AND DISCUSSION

Cloning and sequencing the ns protein genes of SFV

The 3′ third of the 42S RNA genome of SFV, the 26S RNA region, coding for the four viral structural proteins has been cloned and sequenced earlier (23, 24). A restriction fragment from the 5′ end of this cDNA clone (kindly provided by dr. H. Garoff) was used as a primer to synthesize cDNA for the ns protein coding region. The cDNA was cloned into pBR322 by dC:dG tailing. Restriction fragments from two clones obtained, pKTH310 and pKTH320 (Fig. 1), were used as primers for further cDNA synthesis. After two cloning steps about 85 % of the ns protein coding region was covered in three blocks. The restriction enzyme map data of these cDNA clones was the basis for the third cloning step. The two gaps were flanked by SalI sites, thus to obtain clones for these regions the ds cDNA was cleaved with SalI and cloned into the SalI site of pBR322. The ns protein coding region was covered in nine overlapping clones (Fig. 1) except the 250 nucleotides from the 5' end, which has been determined earlier by primer extension of viral RNA (37) and the 22 5' terminal nucleotides by direct RNA sequencing of the antigenome (42). To avoid cloning artefacts the colinearity of the cDNA clones with 42S RNA was determined with S1-nuclease (33). The nucleotide sequence of the cDNA clones was determined by the Maxam and Gilbert (28) and Sanger dideoxy (34) methods. The nucleotide sequence data was analyzed by the computer programms of Staden (43) and Peltola (44).

Complete nucleotide sequence of the nonstructural region

The nucleotide sequence of the ns protein genes and the deduced amino acid sequences are shown in Figure 2. The total length of the 42S RNA is 11442 nucleotides plus the poly(A) tail, which is approximately 80-90 nucleotides in length (5, 6). The base composition of the 42S RNA is 26.7 %A, 20.1 % U, 27.0 % G and 26.2 % C. The molecular weight of the 42S RNA is 3.95x10^6 daltons in the sodium form, without the poly(A) tail. There is an open reading frame in the ns region starting with AUG at nucleotide 86 and ending at a termination codon UAA at nucleotides 7379-7381, coding for a 2431 amino acids long polyprotein. The two other reading frames are blocked with multiple stop codons, 118 and 132, respectively. The total length of the 26S RNA, mRNA for the structural proteins, is 4074 nucleotides, from which the 5' noncoding region is 51 nucleotides, giving an overlap of 13 nucleotides with the ns coding region (45). Translation of the structural polyprotein starts 38 nucleotides from the stop codon of the ns region in a different reading frame from that used for the ns proteins (45).

Nucleotide sequence between the genes of nsP3 and nsP4

The nucleotide sequence of the ns protein genes of Sindbis and Middelburg viruses revealed an opal stop codon UGA six amino acids upstream from the NH_2-terminal amino acid of nsP4 (1).In the sequence of SFV at the same position there is an arginine codon CGA, showing a change from U to C when compared to Sindbis and Middelburg. The sequence downstream from this codon is highly conserved especially between SFV and Middelburg, but shows little

```
                                                                                     nsP1
                                                                                      M  A  A  K  V  H  V  D  I  E  A  D
ATGGCGGATGTGTGACATACACGACGCCAAAAGATTTTGTTCCAGCTCCTGCCACCTCGGCTACGCGAGAGATTAACCACCCACGATGGCCGCCAAAGTGCATGTTGATATTGAGGCTGA
        10        20        30        40        50        60        70        80        90       100       110       120

  S  P  F  I  K  S  L  Q  K  A  F  P  S  F  E  V  E  S  L  Q  V  T  P  N  D  H  A  N  A  R  A  F  S  H  L  A  T  K  L  I
CAGCCCATTCATCAAGTCTTTGCAGAAGGCATTTCCGTCGTTCGAGGTGGAGTCATTGCAGGTCACACCAAATGACCATGCAAATGCCAGAGCATTTTCGCACCTGGCTACCAAATTGAT
       130       140       150       160       170       180       190       200       210       220       230       240

  E  Q  E  T  D  K  D  T  L  I  L  D  I  G  S  A  P  S  R  R  M  M  S  T  H  K  Y  H  C  V  C  P  M  R  S  A  E  D  P  E
CGAGCAGGAGACTGACAAAGACACACTCATCTTGGATATCGGCAGTGCGCCTTCCAGGAGAATGATGTCTACGCACAAATACCACTGCGTATGCCCTATGCGCAGCGCAGAAGACCCCGA
       250       260       270       280       290       300       310       320       330       340       350       360

  R  L  D  S  Y  A  K  K  L  A  A  A  S  G  K  V  L  D  R  E  I  A  G  K  I  T  D  L  Q  T  V  M  A  T  P  D  A  E  S  P
AAGGCTCGATAGCTACGCAAAGAAACTGGCAGCGGCCTCCGGGAAGGTGCTGGATAGAGAGATCGCAGGAAAAATCACCGACCTGCAGACCGTCATGGCTACGCCAGACGCTGAATCTCC
       370       380       390       400       410       420       430       440       450       460       470       480

  T  F  C  L  H  T  D  V  T  C  R  T  A  A  E  V  A  V  Y  Q  D  V  Y  A  V  H  A  P  T  S  L  Y  H  Q  A  M  K  G  V  R
TACCTTTTGCCTGCATACAGACGTCACGTGTCGTACGGCAGCCGAAGTGGCCGTATACCAGGACGTGTATGCTGTACATGCACCAACATCGCTGTACCATCAGGCGATGAAAGGTGTCAG
       490       500       510       520       530       540       550       560       570       580       590       600

  T  A  Y  W  I  G  F  D  T  T  P  F  M  F  D  A  L  A  G  A  Y  P  T  Y  A  T  N  W  A  D  E  Q  V  L  Q  A  R  N  I  G
AACGGCGTATTGGATTGGGTTTGACACCACCCCGTTTATGTTTGACGCGCTAGCAGGCGCGTATCCAACCTACGCCACAAACTGGGCCGACGAGCAGGTGTTACAGGCCAGGAACATAGG
       610       620       630       640       650       660       670       680       690       700       710       720

  L  C  A  A  S  L  T  E  G  R  L  G  K  L  S  I  L  R  K  K  Q  L  K  P  C  D  T  V  M  F  S  V  G  S  T  L  Y  T  E  S
ACTGTGTGCAGCATCCTTGACTGAGGGAAGACTCGGCAAACTGTCCATTCTCCGCAAGAAGCAATTGAAACCTTGCGACACAGTCATGTTCTCGGTAGGATCTACATTGTACACTGAGAG
       730       740       750       760       770       780       790       800       810       820       830       840

  R  K  L  L  R  S  W  H  L  P  S  V  F  H  L  K  G  K  Q  S  F  T  C  R  C  D  T  I  V  S  C  E  G  Y  V  V  K  K  I  T
CAGAAAGCTACTGAGGAGCTGGCACTTACCCTCCGTATTCCACCTGAAAGGTAAACAATCCTTTACCTGTAGGTGCGATACCATCGTATCATGTGAAGGGTACGTAGTTAAGAAAATCAC
       850       860       870       880       890       900       910       920       930       940       950       960

  M  C  P  G  L  Y  G  K  T  V  G  Y  A  V  T  Y  H  A  E  G  F  L  V  C  K  T  T  D  T  V  K  G  E  R  V  S  F  P  V  C
TATGTGCCCCGGCCTGTACGGTAAAACGGTAGGGTACGCCGTGACGTATCACGCGGAGGGATTCCTAGTGTGCAAGACCACAGACACTGTCAAAGGAGAAAGAGTCTCATTCCCTGTATG
       970       980       990      1000      1010      1020      1030      1040      1050      1060      1070      1080

  T  Y  V  P  S  T  I  C  D  Q  M  T  G  I  L  A  T  D  V  T  P  E  D  A  Q  K  L  L  V  G  L  N  Q  R  I  V  V  N  G  R
CACCTACGTCCCCTCAACCATCTGTGATCAAATGACTGGCATACTAGCGACCGACGTCACACCGGAGGACGCACAGAAGTTGTTAGTGGGATTGAATCAGAGGATAGTTGTGAACGGAAG
      1090      1100      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200

  T  Q  R  N  T  N  T  M  K  N  Y  L  L  P  I  V  A  V  A  F  S  K  W  A  R  E  Y  K  A  D  L  D  D  E  K  P  L  G  V  R
AACACAGCGAAACACTAACACGATGAAGAACTATCTGCTTCCGATTGTGGCCGTCGCATTTAGCAAGTGGGCGAGGGAATACAAGGCAGACCTTGATGATGAAAAACCTCTGGGTGTCCG
      1210      1220      1230      1240      1250      1260      1270      1280      1290      1300      1310      1320

  E  R  S  L  T  C  C  C  L  W  A  F  K  T  R  K  M  H  T  M  Y  K  K  P  D  T  Q  T  I  V  K  V  P  S  E  F  N  S  F  V
AGAGAGGTCACTTACTTGCTGCTGCTTGTGGGCATTTAAAACGAGGAAGATGCACACCATGTACAAGAAACCAGACACCCAGACAATAGTGAAGGTGCCTTCAGAGTTTAACTCGTTCGT
      1330      1340      1350      1360      1370      1380      1390      1400      1410      1420      1430      1440

  I  P  S  L  W  S  T  G  L  A  I  P  V  R  S  R  I  K  M  L  L  A  K  K  T  K  R  E  L  I  P  V  L  D  A  S  S  A  R  D
CATCCCGAGCCTATGGTCTACAGGCCTCGCAATCCCAGTCAGATCACGCATTAAGATGCTTTTGGCCAAGAAGACCAAGCGAGAGTTAATACCTGTTCTCGACGCGTCGTCAGCCAGGGA
      1450      1460      1470      1480      1490      1500      1510      1520      1530      1540      1550      1560

  A  E  Q  E  E  K  E  R  L  E  A  E  L  T  R  E  A  L  P  P  L  V  P  I  A  P  A  E  T  G  V  V  D  V  D  V  E  E  L  E
TGCTGAACAAGAGGAGAAGGAGAGGTTGGAGGCCGAGCTGACTAGAGAAGCCTTACCACCCCTCGTCCCCATCGCGCCGGCGGAGACGGGAGTCGTCGACGTCGACGTTGAAGAACTAGA
      1570      1580      1590      1600      1610      1620      1630      1640      1650      1660      1670      1680
                nsP2
  Y  H  A  G  A  G  V  V  E  T  P  R  S  A  L  K  V  T  A  Q  P  N  D  V  L  L  G  N  Y  V  V  L  S  P  Q  T  V  L  K  S
GTATCACGCAGGTGCAGGGGTCGTGGAAACACCTCGCAGCGCGTTGAAAGTCACCGCACAGCCGAACGACGTACTACTAGGAAATTACGTAGTTCTGTCCCCGCAGACCGTGCTCAAGAG
      1690      1700      1710      1720      1730      1740      1750      1760      1770      1780      1790      1800

  S  K  L  A  P  V  H  P  L  A  E  Q  V  K  I  I  T  H  N  G  R  A  G  G  Y  Q  V  D  G  Y  D  G  R  V  L  L  P  C  G  S
CTCCAAGTTGGCCCCCGTGCACCCTCTAGCAGAGCAGGTGAAAATAATAACACATAACGGGAGGGCCGGCGGTTACCAGGTCGACGGATATGACGGCAGGGTCCTACTACCATGTGGATC
      1810      1820      1830      1840      1850      1860      1870      1880      1890      1900      1910      1920

  A  I  P  V  P  E  F  Q  A  L  S  E  S  A  T  M  V  Y  N  E  R  E  F  V  N  R  K  L  Y  H  I  A  V  H  G  P  S  L  N  T
GGCCATTCCGGTCCCTGAGTTTCAAGCTTTGAGCGAGAGCGCCACTATGGTGTACAACGAAAGGGAGTTCGTCAACAGGAAACTATACCATATTGCCGTTCACGGACCGTCGCTGAACAC
      1930      1940      1950      1960      1970      1980      1990      2000      2010      2020      2030      2040

  D  E  E  N  Y  E  K  V  R  A  E  R  T  D  A  E  Y  V  F  D  V  D  K  K  C  C  V  K  R  E  E  A  S  G  L  V  L  V  G  E
CGACGAGGAGAACTACGAGAAAGTCAGAGCTGAAAGAACTGACGCCGAGTACGTGTTCGACGTAGATAAAAAATGCTGCGTCAAGAGAGAGGAAGCGTCGGGTTTGGTGTTGGTGGGAGA
      2050      2060      2070      2080      2090      2100      2110      2120      2130      2140      2150      2160

  L  T  N  P  P  F  H  E  F  A  Y  E  G  L  K  I  R  P  S  A  P  Y  K  T  T  V  V  G  V  F  G  V  P  G  S  G  K  S  A  I
GCTAACCAACCCCCCGTTCCATGAATTCGCCTACGAAGGGCTGAAGATCAGGCCGTCGGCACCATATAAGACTACAGTAGTAGGAGTCTTTGGGGTTCCGGGATCAGGCAAGTCTGCTAT
      2170      2180      2190      2200      2210      2220      2230      2240      2250      2260      2270      2280

  I  K  S  L  V  T  K  H  D  L  V  T  S  G  K  K  E  N  C  Q  E  I  V  N  D  V  K  K  H  R  G  K  G  T  S  R  E  N  S  D
TATTAAGAGCCTCGTGACCAAACACGATCTGGTCACCAGCGGCAAGAAGGAGAACTGCCAGGAAATAGTTAACGACGTGAAGAAGCACCGCGGGAAGGGGACAAGTAGGGAAAACAGTGA
      2290      2300      2310      2320      2330      2340      2350      2360      2370      2380      2390      2400

  S  I  L  L  N  G  C  R  R  A  V  D  I  L  Y  V  D  E  A  F  A  C  H  S  G  T  L  L  A  L  I  A  L  V  K  P  R  S  K  V
CTCCATCCTGCTAAACGGGTGTCGTCGTGCCGTGGACATCCTATATGTGGACGAGGCTTTCGCTTGCCATTCCGGTACTCTGCTGGCCCTAATTGCTCTTGTTAAACCTCGGAGCAAAGT
      2410      2420      2430      2440      2450      2460      2470      2480      2490      2500      2510      2520

  V  L  C  G  D  P  K  Q  C  G  F  F  N  M  M  Q  L  K  V  N  F  N  H  N  I  C  T  E  V  C  H  K  S  I  S  R  R  C  T  R
GGTGTTATGCGGAGACCCCAAGCAATGCGGATTCTTCAATATGATGCAGCTTAAGGTGAACTTCAACCACAACATCTGCACTGAAGTATGTCATAAAAGTATATCCAGACGTTGCACGCG
      2530      2540      2550      2560      2570      2580      2590      2600      2610      2620      2630      2640

  P  V  T  A  I  V  S  T  L  H  Y  G  G  K  M  R  T  T  N  P  C  N  K  P  I  I  I  D  T  T  G  Q  T  K  P  K  P  G  D  I
TCCAGTCACGGCCATCGTGTCTACGTTGCACTACGGAGGCAAGATGCGCACGACCAACCCGTGCAACAAACCCATAATCATAGACACCACAGGACAGACCAAGCCCAAGCCAGGAGACAT
      2650      2660      2670      2680      2690      2700      2710      2720      2730      2740      2750      2760
  V  L  T  C  F  R  G  W  A  K  Q  L  Q  L  D  Y  R  G  H  E  V  M  T  A  A  A  S  Q  G  L  T  R  K  G  V  Y  A  V  R  Q
CGTGTTAACATGCTTCCGAGGCTGGGCAAAGCAGCTGCAGTTGGACTACCGTGGACACGAAGTCATGACAGCAGCAGCATCTCAGGGCCTCACCCGCAAAGGGGTATACGCCGTAAGGCA
      2770      2780      2790      2800      2810      2820      2830      2840      2850      2860      2870      2880

  K  V  N  E  N  P  L  Y  A  P  A  S  E  H  V  N  V  L  L  T  R  T  E  D  R  L  V  W  K  T  L  A  G  D  P  W  I  K  V  L
GAAGGTGAATGAAAATCCCTTGTATGCCCCTGCGTCGGAGCACGTGAATGTACTGCTGACGCGCACTGAGGATAGGCTGGTGTGGAAAACGCTGGCCGGCGATCCCTGGATTAAGGTCCT
      2890      2900      2910      2920      2930      2940      2950      2960      2970      2980      2990      3000

  S  N  I  P  Q  G  N  F  T  A  T  L  E  E  W  Q  E  E  H  D  K  I  M  K  V  I  E  G  P  A  A  P  V  D  A  F  Q  N  K  A
ATCAAACATTCCACAGGGTAACTTTACGGCCACATTGGAAGAATGGCAAGAAGAACACGACAAAATAATGAAGGTGATTGAAGGACCGGCTGCGCCTGTGGACGCGTTCCAGAACAAAGC
      3010      3020      3030      3040      3050      3060      3070      3080      3090      3100      3110      3120

  N  V  C  W  A  K  S  L  V  P  V  L  D  T  A  G  I  R  L  T  A  E  E  W  S  T  I  I  T  A  F  K  E  D  R  A  Y  S  P  V
GAACGTGTGTTGGGCGAAAAGCCTGGTGCCTGTCCTGGACACTGCCGGAATCAGATTGACAGCAGAGGAGTGGAGCACCATAATTACAGCATTTAAGGAGGACAGAGCTTACTCTCCAGT
      3130      3140      3150      3160      3170      3180      3190      3200      3210      3220      3230      3240

  V  A  L  N  E  I  C  T  K  Y  Y  G  V  D  L  D  S  G  L  F  S  A  P  K  V  S  L  Y  Y  E  N  N  H  W  D  N  R  P  G  G
GGTGGCCTTGAATGAAATTTGCACCAAGTACTATGGAGTTGACCTGGACAGTGGCCTGTTTTCTGCCCCGAAGGTGTCCCTGTATTACGAGAACAACCACTGGGATAACAGACCTGGTGG
      3250      3260      3270      3280      3290      3300      3310      3320      3330      3340      3350      3360

  R  M  Y  G  F  N  A  A  T  A  A  R  L  E  A  R  H  T  F  L  K  G  Q  W  H  T  G  K  Q  A  V  I  A  E  R  K  I  Q  P  L
AAGGATGTATGGATTCAATGCCGCAACAGCTGCCAGGCTGGAAGCTAGACATACCTTCCTGAAGGGGCAGTGGCATACGGGCAAGCAGGCAGTTATCGCAGAAAGAAAAATCCAACCGCT
      3370      3380      3390      3400      3410      3420      3430      3440      3450      3460      3470      3480

  S  V  L  D  N  V  I  P  I  N  R  R  L  P  H  A  L  V  A  E  Y  K  T  V  K  G  S  R  V  E  W  L  V  N  K  V  R  G  Y  H
TTCTGTGCTGGACAATGTAATTCCTATCAACCGCAGGCTGCCGCACGCCCTGGTGGCTGAGTACAAGACGGTTAAAGGCAGTAGGGTTGAGTGGCTGGTCAATAAAGTAAGAGGGTACCA
      3490      3500      3510      3520      3530      3540      3550      3560      3570      3580      3590      3600

  V  L  L  V  S  E  Y  N  L  A  L  P  R  R  R  V  T  W  L  S  P  L  N  V  T  G  A  D  R  C  Y  D  L  S  L  G  L  P  A  D
CGTCCTGCTGGTGAGTGAGTACAACCTGGCTTTGCCTCGACGCAGGGTCACTTGGTTGTCACCGCTGAATGTCACAGGCGCCGATAGGTGCTACGACCTAAGTTTAGGACTGCCGGCTGA
      3610      3620      3630      3640      3650      3660      3670      3680      3690      3700      3710      3720

  A  G  R  F  D  L  V  F  V  N  I  H  T  E  F  R  I  H  H  Y  Q  Q  C  V  D  H  A  M  K  L  Q  M  L  G  G  D  A  L  R  L
CGCCGGCAGGTTCGACTTGGTCTTTGTGAACATTCACACGGAATTCAGAATCCACCACTACCAGCAGTGTGTCGACCACGCCATGAAGCTGCAGATGCTTGGGGGAGATGCGCTACGACT
      3730      3740      3750      3760      3770      3780      3790      3800      3810      3820      3830      3840

  L  K  P  G  G  I  L  M  R  A  Y  G  Y  A  D  K  I  S  E  A  V  V  S  S  L  S  R  K  F  S  S  A  R  V  L  R  P  D  C  V
GCTAAAACCCGGCGGCATCTTGATGAGAGCTTACGGATACGCCGATAAAATCAGCGAAGCCGTTGTTTCCTCCTTAAGCAGAAAGTTCTCGTCTGCAAGAGTGTTGCGCCCGGATTGTGT
      3850      3860      3870      3880      3890      3900      3910      3920      3930      3940      3950      3960

  T  S  N  T  E  V  F  L  L  F  S  N  F  D  N  G  K  R  P  S  T  L  H  Q  M  N  T  K  L  S  A  V  Y  A  G  E  A  M  H  T
CACCAGCAATACAGAAGTGTTCTTGCTGTTCTCCAACTTTGACAACGGAAAGAGACCCTCTACGCTACACCAGATGAATACCAAGCTGAGTGCCGTGTATGCCGGAGAAGCCATGCACAC
      3970      3980      3990      4000      4010      4020      4030      4040      4050      4060      4070      4080

          nsP3
  A  G  C |A  P  S  Y  R  V  K  R  A  D  I  A  T  C  T  E  A  A  V  V  N  A  A  N  A  R  G  T  V  G  D  G  V  C  R  A  V
GGCCGGGTGTGCACCATCCTACAGAGTTAAGAGAGCAGACATAGCCACGTGCACAGAAGCGGCTGTGGTTAACGCAGCTAACGCCCGTGGAACTGTAGGGGATGGCGTATGCAGGGCCGT
      4090      4100      4110      4120      4130      4140      4150      4160      4170      4180      4190      4200

  A  K  K  W  P  S  A  F  K  G  A  A  T  P  V  G  T  I  K  T  V  M  C  G  S  Y  P  V  I  H  A  V  A  P  N  F  S  A  T  T
GGCGAAGAAATGGCCGTCAGCCTTTAAGGGAGCAGCAACACCAGTGGGCACAATTAAAACAGTCATGTGCGGCTCGTACCCCGTCATCCACGCTGTAGCGCCTAATTTCTCTGCCACGAC
      4210      4220      4230      4240      4250      4260      4270      4280      4290      4300      4310      4320

  E  A  E  G  D  R  E  L  A  A  V  Y  R  A  V  A  A  E  V  N  R  L  S  L  S  S  V  A  I  P  L  L  S  T  G  V  F  S  G  G
TGAAGCGGAAGGGGACCGCGAATTGGCCGCTGTCTACCGGGCAGTGGCCGCCGAAGTAAACAGACTGTCACTGAGCAGCGTAGCCATCCCGCTGCTGTCCACAGGAGTGTTCAGCGGCGG
      4330      4340      4350      4360      4370      4380      4390      4400      4410      4420      4430      4440

  R  D  R  L  Q  Q  S  L  N  H  L  F  T  A  M  D  A  T  D  A  D  V  T  I  Y  C  R  D  K  S  W  E  K  K  I  Q  E  A  I  D
AAGAGATAGGCTGCAGCAATCCCTCAACCATCTATTCACAGCAATGGACGCCACGGACGCTGACGTGACCATCTACTGCAGAGACAAAAGTTGGGAGAAGAAAATCCAGGAAGCCATTGA
      4450      4460      4470      4480      4490      4500      4510      4520      4530      4540      4550      4560

  M  R  T  A  V  E  L  L  N  D  D  V  E  L  T  T  D  L  V  R  V  H  P  D  S  S  L  V  G  R  K  G  Y  S  T  T  D  G  S  L
CATGAGGACGGCTGTGGAGTTGCTCAATGATGACGTGGAGCTGACCACAGACTTGGTGAGAGTGCACCCGGACAGCAGCCTGGTGGGTCGTAAGGGCTACAGTACCACTGACGGGTCGCT
      4570      4580      4590      4600      4610      4620      4630      4640      4650      4660      4670      4680

  Y  S  Y  F  E  G  T  K  F  N  Q  A  A  I  D  M  A  E  I  L  T  L  W  P  R  L  Q  E  A  N  E  R  I  C  L  Y  A  L  G  E
GTACTCGTACTTTGAAGGTACGAAATTCAACCAGGCTGCTATTGATATGGCAGAGATACTGACGTTGTGGCCCAGACTGCAGGAGGCAAACGAACGGATATGCCTATACGCGCTGGGCGA
      4690      4700      4710      4720      4730      4740      4750      4760      4770      4780      4790      4800

  T  M  D  N  I  G  S  K  C  P  V  N  D  S  D  S  S  T  P  P  R  T  V  P  C  L  C  R  Y  A  M  T  A  E  R  I  A  R  L  R
AACAATGGACAACATCGGATCCAAATGTCCGGTGAACGATTCCGATTCATCAACACCTCCCAGGACAGTGCCCTGCCTGTGCCGCTACGCAATGACAGCAGAACGGATCGCCCGCCTTAG
      4810      4820      4830      4840      4850      4860      4870      4880      4890      4900      4910      4920

  S  H  Q  V  K  S  M  V  V  C  S  S  F  P  L  P  K  Y  H  V  D  G  V  Q  K  V  K  C  E  K  V  L  L  F  D  P  T  V  P  S
GTCACACCAAGTTAAAAGCATGGTGGTTTGCTCATCTTTTCCCCTCCCGAAATACCATGTAGATGGGGTGCAGAAGGTAAAGTGCGAGAAGGTTCTCCTGTTCGACCCGACGGTACCTTC
      4930      4940      4950      4960      4970      4980      4990      5000      5010      5020      5030      5040

  V  V  S  P  R  K  Y  A  A  S  T  T  D  H  S  D  R  S  L  R  G  F  D  L  D  W  T  T  D  S  S  S  T  A  S  D  T  M  S  L
AGTGGTTAGTCCGCGGAAGTATGCCGCATCTACGACGGACCACTCAGATCGGTCGTTACGAGGGTTTGACTTGGACTGGACCACCGACTCGTCTTCCACTGCCAGCGATACCATGTCGCT
      5050      5060      5070      5080      5090      5100      5110      5120      5130      5140      5150      5160

  P  S  L  Q  S  C  D  I  D  S  I  Y  E  P  M  A  P  I  V  V  T  A  D  V  H  P  E  P  A  G  I  A  D  L  A  A  D  V  H  P
ACCCAGTTTGCAGTCGTGTGACATCGACTCGATCTACGAGCCAATGGCTCCCATAGTAGTGACGGCTGACGTACACCCTGAACCCGCAGGCATCGCGGACCTGGCGGCAGATGTGCACCC
      5170      5180      5190      5200      5210      5220      5230      5240      5250      5260      5270      5280

  E  P  A  D  H  V  D  L  E  N  P  I  P  P  P  R  P  K  R  A  A  Y  L  A  S  R  A  A  E  R  P  V  P  A  P  R  K  P  T  P
TGAACCCGCAGACCATGTGGACCTCGAGAACCCGATTCCTCCACCGCGCCCGAAGAGAGCTGCATACCTTGCCTCCCGCGCGGCGGAGCGACCGGTGCCGGCGCCGAGAAAGCCGACGCC
      5290      5300      5310      5320      5330      5340      5350      5360      5370      5380      5390      5400

  A  P  R  T  A  F  R  N  K  L  P  L  T  F  G  D  F  D  E  H  E  V  D  A  L  A  S  G  I  T  F  G  D  F  D  D  V  L  R  L
TGCCCCAAGGACTGCGTTTAGGAACAAGCTGCCTTTGACGTTCGGCGACTTTGACGAGCACGAGGTCGATGCGTTGGCCTCCGGGATTACTTTCGGAGACTTCGACGACGTCCTGCGACT
      5410      5420      5430      5440      5450      5460      5470      5480      5490      5500      5510      5520
```

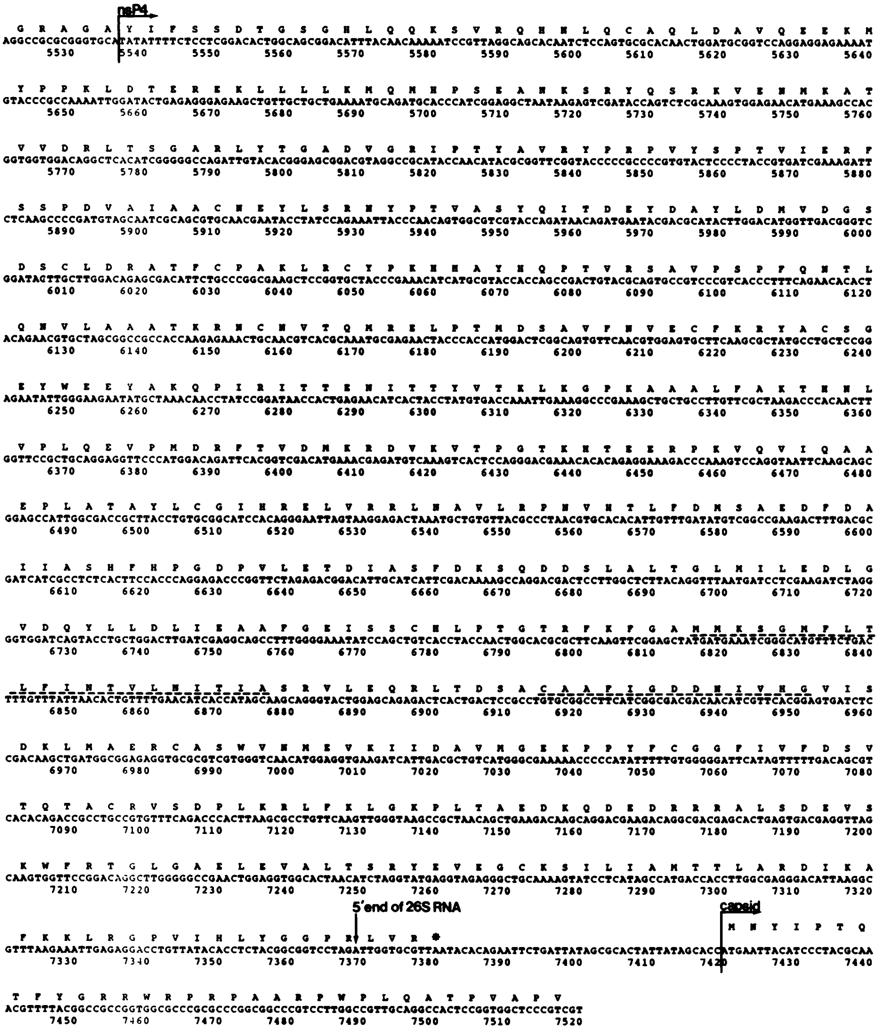

Figure 2. Nucleotide sequence and deduced amino acid sequences of the ns protein genes. Nucleotide sequence is shown in the DNA form. The proteolytic processing sites are marked with arrows. The amino acid repeats in the carboxyterminus of nsP3 are underlined and the region of nsP4 homologous to putative RNA polymerases of other RNA viruses is underlined with a broken line. The 5' end of the 26S RNA and the translation start point of the capsid protein are marked with arrows (45, 23).

homology upstream (Fig. 3a). To rule out the possibility of a cloning artefact at this region, the DNA sequence of the cDNA clone was confirmed by direct RNA sequencing from a specific

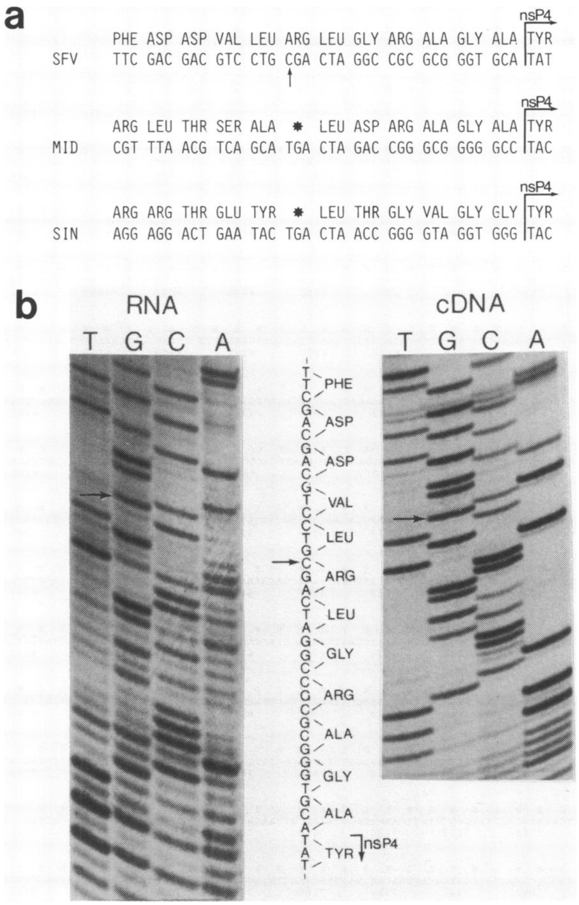
a
nsP4
PHE ASP ASP VAL LEU ARG LEU GLY ARG ALA GLY ALA TYR
SFV TTC GAC GAC GTC CTG CGA CTA GGC CGC GCG GGT GCA TAT
nsP4
ARG LEU THR SER ALA * LEU ASP ARG ALA GLY ALA TYR
MID CGT TTA ACG TCA GCA TGA CTA GAC CGG GCG GGG GCC TAC
nsP4
ARG ARG THR GLU TYR * LEU THR GLY VAL GLY GLY TYR
SIN AGG AGG ACT GAA TAC TGA CTA ACC GGG GTA GGT GGG TAC
b
RNA
T G C A
cDNA
T G C A
T T C G A C G A C G T C C T G C G A C T A G G C C G C G C G G T G C A T A T
PHE
ASP
ASP
VAL
LEU
ARG
LEU
GLY
ARG
ALA
GLY
ALA
TYR
nsP4

Figure 3. a. Comparison of the nucleotide sequences (written in DNA form) of SFV, Middelburg (MID) and Sindbis (SIN) viruses between the genes of nsP3 and nsP4. The opal stop codon (TGA) of Middelburg and Sindbis viruses is marked with a star and the arginine codon (CGA) at the same position in the genome of SFV with an arrow.

b. Sequence analysis of the SFV RNA and cDNA clone from the same region. Both gels show the complementary sequence of the coding strand written in the middle. The C residue in the arginine codon is marked with an arrow.

oligonucleotide primer, which located about 40 nucleotides downstream from the starting point of nsP4. The result of the direct RNA sequencing was identical with the sequence obtained from the cDNA clone, although there was background at some positions, showing that the nucleotide at this position in the genome of SFV is C (Fig. 3b). This finding is compatible with earlier results obtained from the translation studies of the ns proteins of SFV (14, 15, 21, 46), which have shown that nsP4 (ns72) can clearly be detected in cells infected with ts-1 mutant, an overproducer of the ns proteins, and also with the wild type SFV. Furthermore, Keränen and Ruohonen (15) identified the carboxyterminal translation product of SFV in vivo using 30 sec pulses of ^{35}S-methionine. The only carboxyterminal ns protein detected was nsP4 together with its immediate precursor ns135. The existance of nsP4 in Sindbis virus infected cells has been shown only by immunoprecipitation with an antibody prepared against a synthetic COOH-terminal peptide of nsP4 (1). The different expression level of the nsP4 between these virus is, however, surprising, because it is the most conserved ns protein between SFV and Sindbis, indicating an important role for it in the RNA replication.

Deduced amino acid sequences

For the localization of the individual ns protein genes in this nucleotide sequence the partial aminoterminal amino acid sequences of radiolabeled nsP2, nsP3, nsP4 and ns135 were determined (22, Kalkkinen et al. to be published). The NH_2-terminus of nsPl and its precursor ns155 is blocked, but the initiation dipeptide for the ns polyprotein has been shown to be met-ala (47). The AUG codon at position 86-88 is followed by an alanine codon and the only open reading frame in the ns region starts at this point, thus it is quite obvious that nsPl starts at this position. It should, however, be noted that the initiation codon is preceded by two other AUG:s at positions 1 and 8. The proteolytic processing sites yielding the obtained NH_2-termini are marked with arrows in Figure 2. The proteolytic cleavages in the ns polyprotein of SFV occur between alanine-glycine (AG), cysteine-alanine (CA) and alanine-tyrosine (AY), alanine and glycine are preceding all the cleavage sites. The details of the processing steps are unknown, but based on the accumulation of cleavage intermediate (ns220) and the unprocessed polyprotein by certain ts-mutants it has been suggested that one or more of the ns proteins could be involved in this processing (21).

The molecular weights of the ns proteins calculated from the amino

Table 1. Molecular weights of the nonstructural proteins of SFV

	nsP1	nsP2	nsP3	nsP4
Determined by polyacrylamide gel electrophoresis (Keränen & Ruohonen 1983)	64000	86000	61000	68000
Determined from the nucleotide sequence	59600	88500	52100	68900

acid compositions agree quite well with those determined by gel electrophoresis (Table 1), except in the case of nsP3, which has almost 10000 daltons lower molecular weight than estimated earlier. The carboxyterminus of nsP3 has some interesting features. It has a high proline content and it contains amino acid repeats; two identical octapeptides (ADVHPEPA), tetrapeptides (PAPR) and hexapeptides (TFGDFD), these are underlined in Figure 2.

Codon usage

The codon usage of the ns region is shown in Table 2 and compared to that used in Sindbis virus (13). In eukaryotic mRNAs and in some eukaryotic viruses, e.g. poliovirus and VSV, there is a low incidence for CpG dinucleotide codons for serine, proline, threonine and alanine (48, 49, 50). The codon usage in the

Table 2. Codon usage in the nonstructural region of SFV and Sindbis virus

		SFV	SIN			SFV	SIN			SFV	SIN			SFV	SIN
PHE	UUU	32	42	SER	UCU	20	20	TYR	UAU	22	32	CYS	UGU	17	10
	UUC	47	46		UCC	32	24		UAC	60	47		UGC	44	48
LEU	UUA	14	13		UCA	23	28	TERM	UAA	1	0	TERM	UGA	0	1
	UUG	56	39		UCG	29	36		UAG	0	1	TRP	UGG	23	22
LEU	CUU	12	35	PRO	CCU	31	28	HIS	CAU	18	31	ARG	CGU	10	20
	CUC	22	24		CCC	32	32		CAC	45	31		CGC	25	27
	CUA	35	33		CCA	29	50	GLN	CAA	16	37		CGA	14	10
	CUG	76	62		CCG	44	45		CAG	53	53		CGG	10	13
ILE	AUU	28	42	THR	ACU	33	32	ASN	AAU	21	35	SER	AGU	18	21
	AUC	50	51		ACC	49	55		AAC	59	62		AGC	33	31
	AUA	25	31		ACA	49	55	LYS	AAA	66	78	ARG	AGA	45	45
MET	AUG	53	57		ACG	41	33		AAG	80	88		AGG	43	25
VAL	GUU	34	34	ALA	GCU	42	38	ASP	GAU	38	50	GLY	GGU	16	25
	GUC	53	50		GCC	76	76		GAC	105	78		GGC	41	32
	GUA	37	57		GCA	60	63	GLU	GAA	63	88		GGA	46	51
	GUG	84	51		GCG	50	43		GAG	73	75		GGG	29	22

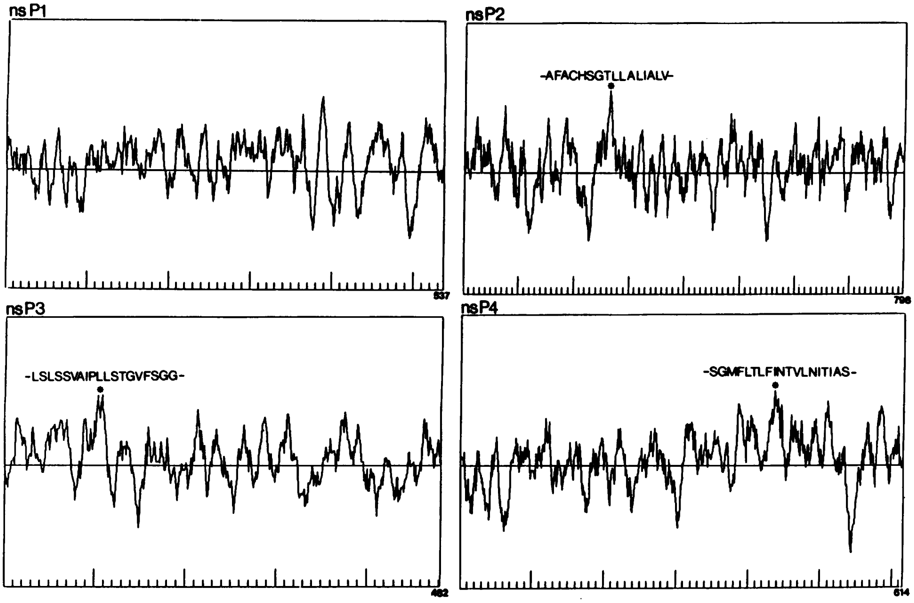

Figure 4. Hydrophobicity plots of the deduced ns proteins of SFV according to Kyte and Doolittle (55) with a search length of seven amino acids. The longest hydrophobic sequences are indicated in one letter code, the scale of each plot shows the length of the protein in amino acids.

structural regions of SFV, Sindbis and Ross River viruses (23, 24, 35, 51) and the ns regions of SFV and Sindbis virus (Table 2) does not show a low CpG incidence, which could reflect the ability of alphaviruses to replicate also in invertebrate hosts. Between SFV and Sindbis virus there is a difference in the codon usage profiles of some amino acids in the ns regions, e.g. in the codons for leucine, valine, tyrosine, histidine and asparagine SFV prefers G or C at the third position in the codon. It is interesting that although the amino acid composition of the ns proteins is extensively conserved, SFV has adapted to a codon usage, in which the codon-anticodon interaction for some amino acids is stronger. Codon usage between different ns proteins and between the structural and the ns regions of SFV are similar.

Hydrophobic features of the ns proteins

The alphavirus replication complex is known to be associated with intracellular membranes (52, 53). NsP1 (ns70), nsP2(ns86) and nsP4 (ns72) are the proteins that are found from the membrane solubilized replication complex (54). The hydrophobicity plots of the ns proteins according to Kyte and Doolittle (55) are shown in Figure 4.

The carboxyterminus of nsP1 has hydrophobic regions separated by hydrophilic residues, nsP2 has hydrophobic peak around residue 260 and nsP3 contains a stretch of 19 uncharged amino acids starting at residue 99. The carboxyterminus of nsP4 contains a gly-asp-asp (GDD) triplet (nt 6929-6937) surrounded by hydrophobic sequences (underlined with a broken line in Fig. 2). These sequences have been shown to be conserved extensively when poliovirus RNA polymerase sequence was compared with many other putative RNA polymerase coding regions of positive strand RNA viruses of eukaryotes and plants and bacterial viruses (56), representing perhaps a functional domain in the polymerase complex. Similar hydrophobic features are also found in the ns proteins of Sindbis virus (13).

A detailed analysis of the conserved sequences of the ns proteins of SFV and Sindbis virus should lead to some predictions about the functions of these proteins; possible protease function, polymerase function and the regulatory factors involved in the RNA replication event.

ACKNOWLEDGEMENTS

I wish to thank Anna-Liisa Ruuska, Meeri Lappalainen and Ritva Rajala for excellent technical assistance and Kirsti Tuominen for secretarial help, and Drs. Hans Söderlund, Sirkka Keränen and Leevi Kääriäinen for valuable advice and critical reading of the manuscrip

The financial support of Neste Foundation is gratefully acknowledged.

REFERENCES

1. Strauss, E.G., Rice, C.M. and Strauss, J.H. (1983) Proc. Natl. Acad. Sci. USA 80, 5271-5275.
2. Kääriäinen, L. and Söderlund, H. (1978) Curr. top. Microbiol. Immunol. 82, 15-69.
3. Garoff, H., Kondor-Koch, C. and Riedel. H. (1982) Curr. Top. Microbiol. Immunol. 92, 1-49.
4. Pettersson, R.F., Söderlund, H. and Kääriäinen, L. (1980) Eur. J. Biochem. 105, 435-443.
5. Glegg, J.C.S. and Kennedy, S.I.T. (1974). J. Gen. Virol. 22, 331-345.
6. Sawicki, P.L. and Gomatos. P.J. (1976) J. Virol. 20, 446-464.
7. Burton, C.J. and Kennedy, S.I.T. (1975) J. Gen. Virol. 28, 111-127.

8. Sawicki, D.L., Sawicki, S.G., Keränen, S. and Kääriäinen, L. 81981) J. Virol. 39, 348-358.
9. Wengler, G. and Wengler, G. (1976) Virology 73, 190-199.
10. Burge, B.W. and Pfefferkorn, E.R. (1966) Virology 30, 201-213.
11. Burge, B.W. and Pfefferkorn, E.R. (1966) Virology 30, 214-223.
12. Strauss, E.G., Lenches, E.M. and Strauss, J.H. (1976) Virology 74, 154-168.
13. Strauss, E.G., Rice, C.M. and Strauss, J.H. (1984) Virology 133, 92-110.
14. Lachmi, B. and Kääriäinen, L. (1976) Proc. Natl. Acad. Sci. USA 73, 1936-1940.
15. Keränen, S. and Ruohonen, L. (1983) J. Virol. 47, 505-515.
16. Keränen, S. and Kääriäinen, L. (1975) J. Virol. 16, 388-396.
17. Lachmi, B., Glanville, N., Keränen, S. and Kääriäinen, L. (1975). J. Virol. 16, 1615-1629.
18. Glanville, N., Lachmi, B., Smith, A. and Kääriäinen, L. (1978) Biochem. Biophys. Acta 518, 497506.
19. Kääriäinen, L., Sawicki, D. and Gomatos, P.J. (1978) J. Gen. Virol. 39, 463-473.
20. Lehtovaara, P., Ulmanen, I., Kääriäinen, L., Keränen, S. and Philipson, L. (1980) Eur. J. Biochem. 112, 461-468.
21. Keränen, S. and Kääriäinen, L. (1979) J. Virol. 32, 19-29.
22. Kalkkinen, N., Laaksonen, M., Söderlund, H. and Jörnvall, H. (1981) Virology 113, 188-195.
23. Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. and Delius, H. (1980) Proc. Natl. Acad. Sci. USA 77, 6376-6380.
24. Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. and Delius, H. (1980) Nature 288, 236-241.
25. Rice, M.C. and Strauss, J.H. (1981) Proc. Natl. Acad. Sci. USA 78, 2062-2066.
26. Tuomi, K., Kääriäinen, L. and Söderlund, H. (1975) Nucl. Acids Res. 2, 555-565.
27. Jeppesen, P.G.N. (1980) Methods in Enzymology 65, 305-319.
28. Maxam, A.M. and Gilbert, W. (1980) Methods in Enzymology 65, 499-560.
29. Gubler, U. and Hoffman, B.J. (1983) Gene 25, 263-269.
30. Chondhung, R., Jay, E. and Wu, R. (1976) Nucl. Acids Res. 3, 101-116.
31. Mandel. M. and Higa, A. (1970) J. Mol. Biol. 53, 159-162.
32. Birnboim, H.C. and Doly, J. (1979) Nucl. Acids Res. 7, 1513-1523.
33. Söderlund, H., Keränen, S., Lehtovaara, P., Palva, I., Pettersson, R.F. and Kääriäinen, L. (1981) Nucl. Acids Res. 9, 3403-3417.
34. Sanger, F., Nickeln, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
35. Smith, H.O. and Birnstiel. M.L. (1976) Nucl. Acids Res. 3, 2387-2398.
36. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning, Cold Spring Harbor Laboratory
37. Lehtovaara, P., Söderlund, H., Keränen, S., Pettersson, R.F. and Kääriäinen, L. (1982) J. Mol. Biol. 156, 731-748.
38. Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107-110.
39. Messing, J. (1983) Methods in Enzymology 101, 21-78.
40. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA 80, 3963-3965.
41. Smith, A.J.H. (1980) Methods in Enzymology 65, 560-580.

42. Wengler, G., Wengler, G. and Gross, H.J. (1979) Nature 282, 754-756.
43. Staden, R. (1980) Nucl. Acids Res. 8, 3673-3694.
44. Peltola, H., Söderlund, H. and Ukkonen, E. (1984) Nucl. Acids Res. 12, 307-321.
45. Riedel, H., Lehrach, H. and Garoff, H. (1981) J. Virol. 42, 725-729.
46. Lachmi, B. and Kääriäinen, L. (1977) J. Virol. 22, 141-149.
47. Glanville, N., Ranki, M., Morser, J., Kääriäinen, L. and Smith, A. (1976) Proc. Natl. Acad. Sci. USA 73, 3059-3063.
48. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucl. Acids Res. 9, r43-r74.
49. Racaniello, V.R. and Baltimore, D. (1981) Proc. Natl. Acad. Sci. USA 78, 4887-4891.
50. Rose, J.K. and Gallione, C.J. (1981) J. Virol. 39, 519-528.
51. Dalgarno, L., Rice, C.M. and Strauss, J.H. (1983) Virology 129, 170-187.
52. Grimley, P.M., Lewin, J.G., Berezesky, I.K. and Friedman, R.M. (1972) J. Virol. 10, 492-503.
53. Friedman, R.M., Levin, J.G., Grimley, P.M. and Berezesky, I.K. (1972) J. Virol. 10, 504-515.
54. Ranki, M. and Kääriäinen, L. (1979) Virology 98, 298-307.
55. Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105-132.
56. Kamer, G. and Argos, P. (1984) Nucl. Acids Res. 12, 7269-7282.